B

Basics about Stochastic Processes and Time Series

This appendix aims to provide a brief survey of some relevant terms and concepts from a basic course about stochastic processes and (continuous-valued) time series analysis. Many more details can be found in introductory textbooks such as those by Box et al. (2015), Brockwell & Davis (2016), Cryer & Chan (2008), Falk et al. (2012), Shumway & Stoffer (2011) and Wei (2006). For Section B.4.2, Lütkepohl (2005) is also recommended as a comprehensive textbook about multivariate time series. For the Markov chains discussed in Section B.2, the book by Seneta (1983) offers a lot of further detail.

B.1 Stochastic Processes: Basic Terms and Concepts

A (discrete-time) stochastic process (or simply process) is a sequence \((X_t)_{\mathcal{T}}\) of random variables \(X_t: \Omega \to S\) defined on a probability space \((\Omega, \mathfrak{F}, P)\) and with range \(S\) (the state space of the process), where \(\mathcal{T}\) is a discrete and linearly ordered set. For simplicity, we usually choose \(\mathcal{T} = \mathbb{Z} := \{\ldots, -1, 0, 1, \ldots\}\) or \(\mathcal{T} = \mathbb{N}_0 := \{0, 1, \ldots\}\). We distinguish between continuous-valued processes and discrete-valued processes, depending on whether the range \(S\) is a continuous or discrete set, respectively. If the range \(S\) of the random variables is equal to the set \(\mathbb{R}\) of real numbers (or to a connected subset thereof), we refer to \((X_t)_{\mathcal{T}}\) as a real-valued process, while a count process even requires \(S \subseteq \mathbb{N}_0\).

If the event \(\omega \in \Omega\) is realized, this leads to a sequence \((X_t(\omega))_{\mathcal{T}}\), a realization (sample path) of the process. A time series \((x_t)_{\mathcal{T}_0}\), where \(\mathcal{T}_0 \subseteq \mathcal{T}\) has to be a finite set in practice, is then understood as an observable part of such a realization, \((x_t)_{\mathcal{T}_0} \subseteq (X_t(\omega))_{\mathcal{T}}\) for a fixed \(\omega \in \Omega\).

If we now speak about a model for the time series, we indeed refer to a model for the underlying process. In the sense of Kolmogorov’s extension theorem, a stochastic process is uniquely characterized through its finite-dimensional distributions; that is, the joint distributions of any finite selection of random...
variables $X_{t_1}, \ldots, X_{t_n}$ with $t_1 < \ldots < t_n \in T$. Besides these complete distributions, certain moment properties are especially relevant for practice.

**Definition B.1.1 (Moments)** If the process $(X_t)_\mathbb{Z}$ consists of square-integrable real-valued random variables, then

- **mean** $\mu(t) := E[X_t],$
- **variance** $\sigma^2(t) := V[X_t],$
- **autocovariance** (at lag $k$) $\gamma_t(k) := \text{Cov}[X_t, X_{t-k}],$
- **autocorrelation** (at lag $k$) $\rho_t(k) := \frac{\text{Corr}[X_t, X_{t-k}]}{\sigma(t) \sigma(t-k)}$

are (generally) functions of the time $t$, with time lag $k \in \mathbb{Z}$.

We abbreviate the autocorrelation function as ACF. Note that $\sigma^2(t) = \gamma_t(0)$.

**Example B.1.2 (White noise)** The most basic type of process is so-called white noise. $(\epsilon_t)_\mathbb{Z}$ is said to be (strong) white noise if the $\epsilon_t$ are independent and identically distributed (i.i.d.) random variables. In the following, we shall usually consider real-valued and square-integrable white noise; because of the i.i.d.-assumption, the above first and second-order moments (Definition B.1.1) are constant in time; see also the stationarity concepts in Definition B.1.3 below.

In some textbooks, $(\epsilon_t)_\mathbb{Z}$ is only required to consist of square-integrable and uncorrelated random variables, with time-invariant first- and second-order moments; such kind of process is referred to as weak white noise hereafter.

Requiring for the time-invariance of certain stochastic properties of the considered process leads to concept of stationarity.

**Definition B.1.3 (Stationarity)** A process $(X_t)_\mathbb{Z}$ is said to be (strictly) stationary if the joint distributions of any finite selection of random variables $X_{t_1}, \ldots, X_{t_k}$ with $t_1 < \ldots < t_n \in \mathbb{Z}$ are time-invariant; that is, if the joint distribution functions of $(X_{t_1}, \ldots, X_{t_k})$ and $(X_{t_1+k}, \ldots, X_{t_n+k})$ are identical for any $k \in \mathbb{Z}$.

A real-valued process $(X_t)_\mathbb{Z}$ is said to be weakly stationary if it is square-integrable, and if mean and autocovariance are time-invariant; that is, if $\mu(t) = \mu$ and $\gamma_t(k) = \gamma(k)$ for all $t \in \mathbb{Z}$ and for any lag $k \in \mathbb{Z}$, which includes the variance for $k = 0$.

So the white noise from Example B.1.2 is stationary, and if it is also real-valued and square-integrable, then it is also weakly stationary. A weak white noise is weakly stationary by definition.
Since, for a weakly stationary process, mean and autocovariances are constant in time, they can be estimated from a given time series of the process by computing the respective empirical counterparts.

**Definition B.1.4 (Sample moments)** Let \( x_1, \ldots, x_T \) be a time series from a weakly stationary real-valued process \((X_t)_{\mathbb{Z}}\). Then we define the

- **sample mean** \( \bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t \),
- **sample variance** \( s^2 = \frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})^2 \),

and for lags \( k = 0, \ldots, T-1 \), we define the

- **sample autocovariance** \( \hat{\gamma}(k) = \frac{1}{T} \sum_{t=k+1}^{T} (x_t - \bar{x})(x_{t-k} - \bar{x}) \),
- **sample autocorrelation** \( \hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} = \frac{\sum_{t=k+1}^{T} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \).

A plot of the sample autocorrelation function (SACF) \( \hat{\rho}(k) \) against the time lag \( k \) is referred to as a correlogram.

Note that \( \hat{\gamma}(0) = s^2 \). The definitions are extended to negative lags by setting \( \hat{\gamma}(-k) := \hat{\gamma}(k) \) and \( \hat{\rho}(-k) := \hat{\rho}(k) \).

As an example, let us consider the sample mean \( \bar{X} \) as an estimator of \( \mu \). If the underlying process is i.i.d. (white noise), then weak moment conditions suffice to guarantee that \( \sqrt{T} (\bar{X} - \mu) \) is asymptotically normally distributed; see, for example, the Lindeberg–Lévy central limit theorem (CLT). To obtain a similar assertion for serially dependent (but still stationary) processes, the dependence structure of the process has to be analyzed in more detail. The traditional approach is to express the extent of serial dependence in terms of **mixing properties**, but more recently, **weak dependence conditions** have also been proposed as an alternative. The following definition summarizes two important mixing conditions for which appropriate CLTs are available; a survey about further mixing properties and their applications is provided by Bradley (2005). The relationship between weak dependence and mixing in the case of discrete-valued processes is discussed by Doukhan et al. (2012).

**Definition B.1.5 (\( \phi \)-mixing, \( \alpha \)-mixing)** Let \((X_t)_{\mathbb{Z}}\) be a stationary process on the probability space \((\Omega, \mathcal{F}, P)\). It is said to be **\( \phi \)-mixing** if there exists a non-negative sequence \((\phi_n)_{n\in\mathbb{N}}\) of weights with \( \phi_n \to 0 \) for \( n \to \infty \) such that for each \( t \in \mathbb{Z}, k \in \mathbb{N} \) and all events \( E_1 \in \mathcal{F}(X_t, X_{t-1}, \ldots) \), \( E_2 \in \mathcal{F}(X_{t+k}, X_{t+k+1}, \ldots) \), the following inequality holds:

\[
|P(E_1 \cap E_2) - P(E_1)P(E_2)| \leq \phi_k P(E_1).
\]
It is called \(\alpha\)-mixing instead (also strongly mixing), with sequence of weights \((\alpha_n)_n\), if the weaker requirement
\[
|P(E_1 \cap E_2) - P(E_1) P(E_2)| \leq \alpha_k
\]
holds.

Both types of mixing process can be understood as a process in which the distant future is approximately independent of the past and present. A useful CLT for \(\phi\)-mixing processes is given, among others, on p. 200 in Billingsley (1999), and a CLT for \(\alpha\)-mixing processes is presented in Theorem 1.7 of Ibragimov (1962). Both theorems make additional requirements on the speed of convergence of the weights towards 0; these requirements are satisfied if the weights can be shown to decrease geometrically quickly.

Examples of non-stationary processes include processes exhibiting trend or seasonality. A particularly simple example of a non-stationary process is the random walk.

**Example B.1.6 (Random walk)** If \((\epsilon_t)_n\) is real-valued white noise, then the process \((X_t)_n\) defined by
\[
X_1 = \epsilon_1, \quad X_t = X_{t-1} + \epsilon_t \quad \text{for } t \geq 2,
\]
is referred to as a random walk. Explicitly, it holds that
\[
X_t = \sum_{s=1}^{t} \epsilon_s.
\]
Hence, if the white noise is even square-integrable with mean \(\mu_{\epsilon}\) and variance \(\sigma^2_{\epsilon}\), we obtain that
\[
\mu(t) = \mu_{\epsilon} t, \quad \sigma^2(t) = \sigma^2_{\epsilon} t,
\]
\[
\gamma_{\epsilon}(k) = \sigma^2_{\epsilon} (t-k)\quad \rho_{\epsilon}(k) = \sqrt{1 - \frac{k}{t}} \quad \text{for } k \geq 0.
\]
If \(\mu_{\epsilon} \neq 0\), then the mean is not constant in time; in such a case, one also speaks of a random walk with drift.

By construction, the random walk at future time \(t+1\) remembers its present value \(X_t\) and then adds a random disturbance \(\epsilon_{t+1}\); further past values \(X_{t-1}, X_{t-2}, \ldots\) are without influence on \(X_{t+1}\) if the value of \(X_t\) is given. A process with such a “memory of length 1” is called a Markov chain.

**Definition B.1.7 (Markov process)** A process \((X_t)_Z\) with state space \(S\) is said to be a \(p\)th-order Markov process with \(p \in \mathbb{N}\) if for all \(t \in \mathbb{Z}\) and for each measurable set \(A \subseteq S\),
\[
P(X_t \in A \mid X_{t-1} = x_{t-1}, \ldots) = P(X_t \in A \mid X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p})
\]
holds for all \(x_i \in S\). In the case \(p = 1\), \((X_t)_Z\) is commonly called a Markov chain.
In this book, only the case of a discrete state space is of relevance. Then, the Markov property of Definition B.1.7 simplifies to

\[
P(X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots) = P(X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p}) ,
\]

(B.1)

which has to hold for all \(x_t \in S\). In fact, for theoretical analysis, it suffices to concentrate on the first-order case, since any pth-order Markov process with state space \(S\) can be transformed into a first-order Markov process with state space \(S^p\) by considering the vector-valued process \((X_t)_{t \geq 0}\) with \(X_t := (X_t, \ldots, X_{t-p+1})^T\). The next section summarizes important properties of such discrete-valued Markov chains.

### B.2 Discrete-Valued Markov Chains

In this section, the process \((X_t)_{t \geq 0}\) is assumed to be a (first-order) Markov chain with state space \(S\). \(S\) is either finite or countably infinite; if \(S\) is not even ordinal, then it is assumed that its values are at least arranged in a certain lexicographical order to simplify notations: \(S = \{s_0, s_1, \ldots\}\). Here, we summarize some basic facts about such discrete-valued Markov chains, but much more information is provided in, for example, the book by Seneta (1983) or in Chapter XV of Feller (1968).

#### B.2.1 Basic Terms and Concepts

From now on, we shall always assume that the Markov chain \((X_t)_{t \geq 0}\) is even (time-)homogeneous; that is, for any \(i, j \in S\), the transition probabilities do not vary with time:

\[
P(X_t = i \mid X_{t-1} = j) = p_{ij} \quad \text{for all } t \in \mathbb{N} .
\]

(B.2)

The corresponding transition matrix \(P = (p_{ij})_{i,j \in S, t \geq 0} \) might be of infinite dimension, depending on the cardinality of \(S\). Similarly, all marginal probabilities at a given time are summarized as a (possibly time-dependent) vector:

\[
p_t := (P(X_t = s_0), P(X_t = s_1), \ldots)^T.
\]

Knowing the marginal distribution at time 0 (initial distribution), the transition matrix allows us to compute all further probabilities recursively via

\[
p_t = P p_{t-1} = \ldots = P^t p_0 .
\]

(B.3)

In particular, the \(h\)-step-ahead transition probabilities \(p_{ij}^{(h)} := P(X_t = i \mid X_{t-h} = j)\) are the entries of the matrix \(P^h\).
Example B.2.1.1 (Classification of states) With the help of the \( h \)-step-ahead transition probabilities \( p_{ij}^{(h)} \), the Markov chain’s states \( S \) can be classified. If, for given \( i, j \in S \), there exists a time lag \( h \) such that \( p_{ij}^{(h)} > 0 \), we say that \( j \) leads to \( i \). If both \( j \) leads to \( i \) and \( i \) leads to \( j \), then \( i \) and \( j \) are said to be communicating states.

If for the state \( i \in S \), we find a state \( j \) such that \( i \) leads to \( j \), but \( j \) does not lead to \( i \), then \( i \) is an inessential state. If an essential state \( i \), in turn, leads to a state \( j \), then it necessarily communicates with \( j \).

In view of (B.3), it is obvious that the homogeneous Markov chain \( (X_t)_{t \geq 0} \) becomes stationary (see Definition B.1.3) if the marginal distributions remain fixed in time; that is, if \( p_t = \ldots = p_0 =: p \). So in view of (B.3), we need to look at so-called invariant probability vectors; that is, at probability vectors \( \pi \) satisfying the invariance equation

\[
P \pi = \pi. \tag{B.4}
\]

Then \( (X_t)_{t \geq 0} \) is stationary iff the initial distribution is an invariant vector; that is, the invariant distributions are the possible stationary marginal distributions of \( (X_t)_{t \geq 0} \). The essential question is as follows:

Which conditions must be satisfied by the transition matrix \( P \) such that an invariant distribution (uniquely?) exists at all? Possible answers to this question are presented in Section B.2.2 below. Before this, let us briefly discuss the maximum likelihood approach to parameter estimation.

Remark B.2.1.2 (Maximum likelihood and information criteria) A popular approach for parameter estimation of a model for a stochastic process (not necessarily a Markov process) is the maximum likelihood (ML) approach. The intuition behind this approach is to select the parameter values such that the observed time series \( x_1, \ldots, x_T \) becomes most “plausible”. In the discrete-valued case, with the model parameters summarized in the vector \( \theta \), the likelihood function is defined as the probability of observing what has already been observed, as a function of the unknown \( \theta \):

\[
L(\theta) := P(X_T = x_T, \ldots, X_1 = x_1 \mid \theta).
\]

The ML estimate of \( \theta \) is defined to be a value \( \hat{\theta}_{\text{ML}} \) maximizing \( L(\theta) \). In view of computational issues, it is common practice to determine \( \hat{\theta}_{\text{ML}} \) not by maximizing \( L(\theta) \), but by maximizing the logarithmic likelihood function (log-likelihood) instead; that is, \( \ell(\theta) := \ln(L(\theta)) \). We denote the value of the maximized log-likelihood as \( \ell_{\text{max}} \).

In general, it is not clear if the ML estimators (uniquely) exist and if they are consistent. Even if they exist, they can often only be computed numerically by using appropriate optimization routines. The computation of the (log-)likelihood function itself is also sometimes not easy. An exception are
homogeneous Markov chains, for which the joint distributions of segments from \((X_t)_{t=0}^T\) are easily computed,

\[
P(X_r = x_r, \ldots, X_s = x_s) = P(X_r = x_r) \cdot \prod_{i=r+1}^{s} P_{x_i | x_{i-1}}, \quad r < s.
\]

Similar arguments apply to \(p\)-th-order Markov processes, for which we obtain

\[
L(\theta) = P(X_1 = x_1, \ldots, X_p = x_p) \cdot \prod_{t=p+1}^{T} P(x_t | x_{t-1}, \ldots, x_{t-p}),
\]

\[
\ell(\theta) = \ln P(X_1 = x_1, \ldots, X_p = x_p) + \sum_{t=p+1}^{T} \ln p(x_t | x_{t-1}, \ldots, x_{t-p}),
\]

(B.5)

where \(p(x | x_{-1}, \ldots, x_{-p})\) denotes the time-homogeneous \(p\)-th-order conditional probabilities for all \(x, x_{-1}, \ldots, x_{-p} \in S\). If the Markov process is stationary, the initial probability \(P(X_1 = x_1, \ldots, X_p = x_p)\) can be obtained as a function of the conditional probabilities (B.4). To avoid this step, one often simply maximizes the conditional (log-)likelihood,

\[
L(\theta | x_p, \ldots, x_1) = \prod_{t=p+1}^{T} p(x_t | x_{t-1}, \ldots, x_{t-p}),
\]

\[
\ell(\theta | x_p, \ldots, x_1) = \sum_{t=p+1}^{T} \ln p(x_t | x_{t-1}, \ldots, x_{t-p}),
\]

(B.6)

and refers to the resulting estimates as conditional ML estimates \(\hat{\theta}_{CML}\). Results concerning the existence, consistency and asymptotic normality of the (C)ML estimators for discrete-valued Markov chains are available in Part I of the book by Billingsley (1961). Under appropriate conditions (Billingsley, 1961), \(\sqrt{T-p} (\hat{\theta}_{CML} - \theta)\) is asymptotically normally distributed according to \(N(0, \mathbf{I}^{-1}(\theta))\), where \(\mathbf{0}\) denotes the zero vector, and where \(\mathbf{I}(\theta)\) denotes the expected Fisher information per observation. The latter is defined as follows. The negative Hessian of \(\ell(\theta)\) is said to be the observed Fisher information \(\mathbf{J}(\theta)\), which can be written as \(\mathbf{J}(\theta) = \sum_{t=p+1}^{T} J_t(\theta)\) with \(J_t(\theta)\) being the Hessian of \(- \ln p(x_t | x_{t-1}, \ldots, x_{t-p})\); then \(\mathbf{I}(\theta)\) is the expectation of \(\mathbf{J}_t(\theta)\). In practice, the mean observed Fisher information \(\frac{1}{T-p} \mathbf{J}(\theta)\) can be used to approximate \(\mathbf{I}(\theta)\).

Plugging in the obtained estimates \(\hat{\theta}_{CML}\) in \((\mathbf{J}(\theta))^{-1}\) allows us, for instance, to approximate the asymptotic standard errors of the CML estimators.

The ML approach can be applied in combination with so-called information criteria for model selection (if several candidate models are available for the time series). On the one hand, it is desirable to select a model with the largest \(\ell_{\text{max}}\) possible. On the other hand, small models are more easily interpretable and also overfitting has to be avoided. The idea behind information criteria is to balance the goodness of fit against the model size by adding an appropriate penalty term to \(\ell_{\text{max}}\). Let \(n_{\text{model}}\) abbreviate the number of parameters of the given model, then Akaike’s information criterion (AIC) and the
Bayesian information criterion (BIC) are given by
\[
\text{AIC} = -2 \ell_{\text{max}} + 2 n_{\text{model}}, \quad \text{BIC} = -2 \ell_{\text{max}} + n_{\text{model}} \ln T, \quad (B.7)
\]
respectively. That model producing the lowest value of AIC or BIC, respectively, is selected from among all candidate models. With increasing \(T\), the BIC more strongly penalizes the model size, so it tends to select smaller models than the AIC. If applied to estimate the order \(p\) of a finite Markov process, Katz (1981) showed that only the BIC is consistent while the AIC leads to an overestimation.

AIC and BIC can also be computed based on the CML approach. Since the number of terms in \(\ell(\theta \mid x_p, \ldots, x_1)\) from (B.6) varies with varying \(p\), one may insert the factor \(T / (T - p)\) before \(\ell_{\text{max}}\) in (B.7) to account for this distortion.

### B.2.2 Stationary Markov Chains

Criteria for the stationarity of a Markov chain will depend on the cardinality of \(S\); the existence of an invariant distribution (and hence of a stationary solution to the Markov condition) will be much easier to establish if \(S\) is finite (finite Markov chain).

So let us now introduce, step-by-step, the potentially relevant conditions for the existence of a stationary solution (Feller, 1968; Seneta, 1983). The Markov chain is said to be irreducible if for any \(i, j \in S\), there exists some lag \(h \in \mathbb{N}\) such that \(p_{ij}^{(h)} > 0\). Note that \(h\) may differ for varying \(i, j\). But if there exists a unique \(h \in \mathbb{N}\) such that all \(p_{ij}^{(h)} > 0\), then the Markov chain is said to be primitive.

For a finite Markov chain, the following conclusions can be drawn:

- An irreducible finite Markov chain possesses a unique stationary distribution \(p\), being the unique solution of the invariance equation (B.4).
- A primitive finite Markov chain is even ergodic; that is, the distributions \(p_t\) always converge to the stationary distribution \(p\) for \(t \to \infty\), independent of the initial distribution \(p_0\).

Expressed in terms of the \(h\)-step-ahead transition probabilities \(p_{ij}^{(h)}\), ergodicity means \(p_{ij}^{(h)} \to p_i\) for \(h \to \infty\); that is, the \(h\)-step-ahead forecasting distribution \(P^h\) converges to the marginal distribution \(p^{\top}\), where \(1\) denotes the vector of ones (concerning the rate of convergence, see Remark B.2.2.1 below). This characterization offers a way of numerically computing the stationary distribution (instead of solving the eigenvalue problem (B.4)), by considering \(p_{ij}^{(M)}\) with a sufficiently large \(M\) as an approximation for \(p_i\).

Another approach to ergodicity is to look at the period of the Markov chain. An irreducible Markov chain has a unique period, and this period equals \(d\) if the greatest common divisor of those \(h \in \mathbb{N}\), for which \(p_{ij}^{(h)} > 0\), equals \(d\) (for \(i \in S\) chosen arbitrarily). For \(d = 1\), the Markov chain is said to be aperiodic, a property that may be established by the following criterion: if at least one diagonal element \(p_{ii}\) of \(P\) is positive, then the irreducible Markov chain is aperiodic.
For a finite Markov chain, the following equivalence holds:

- irreducible and aperiodic $\Leftrightarrow$ primitive (and these imply ergodicity).

Finally, in view of the mixing concepts discussed in Definition B.1.5 above, it is worth mentioning that an ergodic finite Markov chain is $\phi$-mixing with geometrically decreasing weights; that is, weights $\phi_n = a \cdot \rho^n$ with $a > 0$ and $0 < \rho < 1$.

**Remark B.2.2.1 (Perron–Frobenius theorem)** As summarized above, a primitive finite Markov chain is even ergodic; that is, $P^h \to p^1^\top$ for $h \to \infty$.

The rate of convergence can be expressed by adapting the Perron–Frobenius theorem. This theorem applies to non-negative matrices $T$ in general; that is, matrices $T$ having non-negative entries $t_{ij} \geq 0$. The transition matrix $P$ is a special case in which the column sums are equal to 1 (a so-called stochastic matrix). The theorem states that for a primitive non-negative matrix $T$, there exists the so-called Perron–Frobenius eigenvalue $\lambda_{PF}$, which satisfies

- $\lambda_{PF} > |\lambda|$ for any other eigenvalue $\lambda \neq \lambda_{PF}$ ("largest eigenvalue");
- $\lambda_{PF}$ takes a value between the minimum and the maximum of all row sums of $T$, and, in the same way, also a value between the minimum and the maximum of all column sums of $T$;
- $\lambda_{PF}$ has geometric and algebraic multiplicity 1;
- with $\lambda_{PF}$ can be associated strictly positive left and right eigenvectors.

In this case, let $v$ and $w$ be corresponding positive left and right eigenvectors such that $v^\top w = 1$. Denote the remaining distinct eigenvalues by $\lambda_2, \ldots, \lambda_r$, where $\lambda_{PF} > |\lambda_2| \geq \ldots \geq |\lambda_r|$ and where $\lambda_2$ ("second largest eigenvalue") has maximal multiplicity among all eigenvalues with the same modulus as $\lambda_2$; denote this multiplicity of $\lambda_2$ by $m_2$. Then the behavior of $T^h$ for growing $h$ is as follows:

$$T^h = \lambda_{PF}^h \cdot wv^\top + O(h^{m_2-1} \cdot |\lambda_2|^h).$$

This general result is now applied to the case of a primitive transition matrix $P$ with (unique) invariant distribution $p$. Since then all column sums are equal to 1, it follows that the Perron–Frobenius eigenvalue equals $\lambda_{PF} = 1$. Furthermore, because of (B.4) – that is, $P \cdot p = p$, and $1^\top \cdot P = 1^\top$ (column sums equal to 1) – any right eigenvector for $\lambda_{PF} = 1$ is a multiple of $p$, and any left eigenvector a multiple of $1$. So it follows that

$$P^h = p1^\top + O(h^{m_2-1} \cdot |\lambda_2|^h);$$

that is, the rate of convergence of $P^h \to p1^\top$ is determined by the absolute value of the second largest eigenvalue, which satisfies $|\lambda_2| < 1$.

Let us conclude this section about discrete-valued Markov chains by looking at the countably infinite case. Here, irreducibility alone does not guarantee
the existence of an invariant distribution. Therefore, it is necessary to look at another characteristic of the states of a Markov chain: the recurrence properties. Let \( r_{jj}(n) \) denote the conditional probability that the Markov chain \((X_t)_{t \in \mathbb{N}_0}\), given that it started in \( X_0 = j \), returns to the state \( j \) for the first time at time \( t = n \); that is,

\[
r_{jj}(n) := P(X_n = j, X_{n-1}, \ldots, X_1 \neq j | X_0 = j).
\]

Summing about all \( n \), we get the probability \( r_{jj} := \sum_{n=1}^{\infty} r_{jj}(n) \) that the Markov chain will once return to the state \( j \). If this probability \( r_{jj} \) equals 1, then \( r_{jj}(1), r_{jj}(2), \ldots \) constitute a valid probability distribution, the recurrence time distribution. The corresponding mean, \( \mu_j = \sum_{n=1}^{\infty} n r_{jj}(n) \), is referred to as the mean recurrence time. Now, the following classification is done:

- The state \( j \) is transient if \( r_{jj} < 1 \); that is, there is a positive probability of never returning to \( j \) at all.
- Otherwise, that is, if \( r_{jj} = 1 \), it is said to be recurrent:
  - it is positive recurrent if its mean recurrence time is finite, \( \mu_j < \infty \),
  - it is null recurrent if its recurrence time is infinite, \( \mu_j = \infty \).

For an irreducible Markov chain, all states are of the same recurrence type, just as they also have a unique period; see above. The following conclusions can be drawn:

- If \((X_t)_{t \in \mathbb{N}_0}\) is an irreducible, aperiodic and positive recurrent\(^1\) Markov chain, then it possesses a unique stationary distribution, and it is also ergodic.
- If \((X_t)_{t \in \mathbb{N}_0}\) is an irreducible and aperiodic Markov chain having an invariant distribution \( p \), then this distribution is unique, and \((X_t)_{t \in \mathbb{N}_0}\) is also positive recurrent and hence ergodic, where \( p_j = 1/\mu_j \) for all \( j \in S \).

Finally, such a countably infinite Markov chain, being irreducible, aperiodic and stationary, is also \( \alpha \)-mixing (Definition B.1.5); but in contrast to the finite case above, there is no further general assertion concerning the speed of convergence of the mixing weights (Bradley, 2005, Section 3).

### B.3 ARMA Models: Definition and Properties

The remaining sections of Appendix B are about common models for continuous-valued time series (Box et al., 2015; Brockwell & Davis, 1991, 2016; Cryer & Chan, 2008; Falk et al., 2012; Shumway & Stoffer, 2011; Wei, 2006). For

\(^1\) Note that a finite Markov chain cannot have a null recurrent state, and it is also impossible that all states are transient. Hence an irreducible finite Markov chain is automatically positive recurrent.
such continuous-valued processes, the class of linear processes is of particular importance.

**Background B.3.1 (Linear processes)** A real-valued sequence \((a_u)_\mathbb{Z}\) of weights, the *filter*, is said to be *absolutely summable* if

\[
\sum_{u=-\infty}^{\infty} |a_u| := \sum_{u=0}^{\infty} |a_u| + \sum_{u=1}^{\infty} |a_{-u}| < \infty.
\]

If \((Z_t)_\mathbb{Z}\) is a weakly stationary process with mean \(\mu_Z\) and autocovariance function \(\gamma_Z(k)\), with \((a_u)_\mathbb{Z}\) being an absolutely summable filter, then the filtered process \((Y_t)_\mathbb{Z}\), defined by

\[
Y_t := \sum_{u=-\infty}^{\infty} a_u Z_{t-u},
\]

exists and is also weakly stationary, where

\[
\mu_Y := E[Y_t] = \mu_Z \sum_{u=-\infty}^{\infty} a_u,
\]

\[
\gamma_Y(k) := \text{Cov}[Y_t, Y_{t-k}] = \sum_{u,v=-\infty}^{\infty} a_u a_v \gamma_Z(k-u+v).
\]

In particular, if the process to be filtered is not only weakly stationary but square-integrable white noise \((\epsilon_t)_\mathbb{Z}\) (the *innovations*), then the *linear process*

\[
X_t := \sum_{u=-\infty}^{\infty} a_u \epsilon_{t-u}
\]

is also stationary with

\[
\mu = \mu_\epsilon \sum_{u=-\infty}^{\infty} a_u, \quad \gamma(k) = \sigma_\epsilon^2 \sum_{u=-\infty}^{\infty} a_u a_{u+k}.
\]  

(B.8)

Note that the autocovariance function is now (nearly) solely defined through the filter \((a_u)_\mathbb{Z}\). Therefore, it is essential to study the properties of such filters \((a_u)_\mathbb{Z}\). For this purpose, consider the Laurent series

\[
A(z) := \sum_{u=-\infty}^{\infty} a_u z^u,
\]

which is referred to as the *characteristic polynomial* of the filter \((a_u)_\mathbb{Z}\). Using the *backshift operator* \(B – \) that is, the operator defined by \(BX_t := X_{t-1} – \) the application of the linear filter is expressed as

\[
X_t = A(B) \epsilon_t.
\]
The successive application of two filters \((a_u)_z\) and \((b_v)_z\) is expressed by multiplying the corresponding characteristic polynomials \(A(z)\) and \(B(z)\). We say that \((a_u)_z\) is invertible (and \((b_v)_z\) is its inverse filter) if \(A(z) B(z) = 1\). So the weights of the inverse filter are obtained by expanding \(A^{-1}(z) = 1/A(z)\).

An absolutely summable filter \((a_u)_z\) is said to be causal if \(a_u = 0\) for all \(u < 0\); that is, if \(A(z) = \sum_{u=0}^{\infty} a_u z^u\) is a power series. The practical meaning of a causal filter is that \(X_t\) does not depend on future innovations. An important example of a causal filter is one having only finitely many non-zero weights,

\[
A_p(z) := 1 - a_1 z - \ldots - a_p z^p. \tag{B.9}
\]

\(A_p(z)\) has an absolutely summable and causal inverse filter iff the \(p\) (possibly complex) roots of \(A_p(z)\) are outside the unit circle; that is, iff they have an absolute value larger than 1.

Now, we are in a position to introduce the family of ARMA models, which is done in three steps. First, we consider the pure moving-average models of order \(q\), where the current observation is defined as a weighted mean of the current observation and \(q\) past innovations. These models date back to works by G. U. Yule and E. E. Slutsky in the 1920s; see Nie & Wu (2013).

**Definition B.3.2 (MA(q) model)** Let \((\epsilon_t)_z\) be square-integrable white noise. Then \((X_t)_z\) defined by

\[
X_t = \epsilon_t - \beta_1 \epsilon_{t-1} - \ldots - \beta_q \epsilon_{t-q} \quad (\beta_q \neq 0)
\]

\[
= \beta(B) \cdot \epsilon_t \quad \text{with} \quad \beta(z) = 1 - \beta_1 z - \ldots - \beta_q z^q
\]

is said to be a moving-average process of order \(q\), abbreviated as MA(q) process.

The case \(q = \infty\) (MA(\(\infty\)) process) just corresponds to the causal linear process defined in Background B.3.1. The MA(q) process is strictly and weakly stationary with

\[
\mu = \mu_\epsilon \left(1 - \beta_\bullet\right), \quad \text{where} \quad \beta_\bullet := \sum_{i=1}^{q} \beta_i, \tag{B.10}
\]

\[
\gamma(k) = \begin{cases} 
0 & \text{for } k > q, \\
\gamma(-k) & \text{for } k < 0, \\
\sigma_\epsilon^2 \sum_{u=0}^{\infty} \beta_u \beta_{u+k} & \text{for } 0 \leq k \leq q, \quad \text{where} \quad \beta_0 := -1;
\end{cases} \tag{B.11}
\]

see (B.8). Note that the autocorrelation function vanishes after lag \(q\); this property is used to identify the model order \(q\) of an MA model: one visually inspects the available correlogram for an (abrupt) drop towards zero; the largest lag with
a (significantly) non-zero autocorrelation is used as an estimate of the model order.

Finally, it should be noted that result (B.9) for linear processes implies that the MA(q) process is invertible with a causal inverse iff the roots of $\beta(z)$ are outside the unit circle. Then $\epsilon_t = \sum_{u=0}^{\infty} \psi_u X_{t-u}$; that is, the innovation at time $t$ can be recovered from the observations available at time $t$, where the coefficients $\psi_u$ are obtained by expanding $\beta^{-1}(z)$ into a power series.

As the second step towards full ARMA models, we consider the autoregressive models of order $p$, where the current observation is a weighted mean of $p$ past observations plus noise. So these processes are generated from their own pasts (hence “auto” regressive; see also the Markov models in Definition B.1.7). These models were established by G. U. Yule and G. T. Walker; see Nie & Wu (2013) for references.

**Definition B.3.3 (AR(p) model)** Let $(\epsilon_t)_\mathbb{Z}$ be square-integrable white noise. Then $(X_t)_\mathbb{Z}$ defined by

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \epsilon_t \quad (\alpha_p \neq 0)$$

$$\Leftrightarrow a(B) \cdot X_t = \epsilon_t \quad \text{with} \quad a(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p$$

is said to be an autoregressive process of order $p$, abbreviated as AR(p) process.

In view of the above result (B.9), a (weakly) stationary and causal solution exists for the AR(p) recursion iff the roots of $\alpha(z)$ are outside the unit circle; then $X_t = \sum_{u=0}^{\infty} \theta_u \epsilon_{t-u}$ (a causal linear process), where the coefficients $\theta_u$ are obtained by expanding $\alpha^{-1}(z)$ into a power series.

If $(X_t)_\mathbb{Z}$ is a weakly stationary and causal AR(p) process, then its mean is computed as

$$\mu = E[X_t] = \mu_\epsilon 1 - \alpha_* \quad \text{where} \quad \alpha_* := \sum_{i=1}^{p} \alpha_i,$$

while the autocorrelation function $\rho(k)$ is obtained by solving the so-called Yule–Walker equations:

$$\rho(k) = \sum_{i=1}^{p} \alpha_i \rho(|k - i|) \quad \text{for} \quad k = 1, 2, \ldots$$

Now, we typically have $\rho(k) \neq 0$ for all $k$ such that $\rho(k)$ cannot be (directly) applied to identify the model order of the autoregressive process, in contrast to the corresponding procedure for moving-average processes, as sketched below formula (B.11). However, transforming the (sample) autocorrelation function into the (sample) partial autocorrelation function first, an analogous procedure is possible.
Theorem B.1 (Partial autocorrelation) Let \( r_k := (\rho(1), \ldots, \rho(k))^\top \) as well as
\[
R_k := (\rho(|i - j|))_{i,j=1,\ldots,k} = \begin{pmatrix}
1 & \rho(1) & \cdots & \rho(k - 1) \\
\rho(1) & 1 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \rho(1) \\
\rho(k - 1) & \cdots & \rho(1) & 1
\end{pmatrix},
\]
assume that \( R_k \) is invertible. Let \( a_k \in \mathbb{R}^k \) be the unique solution of the equation
\[
R_k a_k = r_k,
\]
that is, \( a_k = R_k^{-1} r_k \). (B.14)
Then \( \rho_{\text{part}}(k) := a_{k,k} \) (last component of \( a_k \)) is referred to as the partial autocorrelation at lag \( k \).

If \( (X_t) \) is a weakly stationary and causal AR(p) process, then the partial autocorrelation function (PACF) satisfies
\[
\rho_{\text{part}}(p) = \alpha_p, \quad \rho_{\text{part}}(k) = 0 \quad \text{for } k > p.
\]

Note that the PACF can be computed recursively according to
\[
a_{k+1,k+1} = \frac{\rho(k + 1) - \sum_{i=1}^{k} a_{k,i} \rho(k + 1 - i)}{1 - \sum_{i=1}^{k} a_{k,i} \rho(i)}, \quad a_{k+1,j} = a_{k,j} - a_{k+1,k+1} a_{k,k-j+1} \quad \text{for } j = 1, \ldots, k.
\] (B.15)
The sample version of the PACF (SPACF) is derived in the same way as in (B.14) and (B.15), but with \( \rho(\cdot) \) replaced by the corresponding sample autocorrelations, \( \hat{\rho}(\cdot) \).

As the final step, we combine the AR model with the MA model to obtain the full ARMA model. This combination was suggested by A. M. Walker in 1950, and the embedding into the theory of linear processes (Background B.3.1) was initiated by H. Wold in the late 1930s (Nie & Wu, 2013).

Definition B.3.4 (ARMA(p, q) model) Let \( (\epsilon_t) \) be square-integrable white noise. Then \( (X_t) \), defined by \( (\alpha_p, \beta_q) \neq 0):
\[
X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \epsilon_t - \beta_1 \epsilon_{t-1} - \ldots - \beta_q \epsilon_{t-q}
\]
\[
\Rightarrow \alpha(B) X_t = \beta(B) \epsilon_t \quad \text{with} \quad \alpha(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p
\]
\[
\text{and} \quad \beta(z) = 1 - \beta_1 z - \ldots - \beta_q z^q
\]
is said to be an autoregressive moving-average process of order \((p, q)\), which is abbreviated as ARMA(p, q) process. To keep the order \((p, q)\) minimal, we also require that \( \alpha(z) \) and \( \beta(z) \) have no common roots.

The pure AR and pure MA models are embedded in Definition B.3.4 as the cases \( q = 0 \) and \( p = 0 \), respectively.
With analogous arguments as above (result (B.9)), a (weakly) stationary and causal solution exists for the ARMA(p, q) recursion iff the roots of \( \alpha(z) \) are outside the unit circle; then \( X_t = \sum_{u=0}^{\infty} \theta_u \epsilon_{t-u} \) (a causal linear process), where the coefficients \( \theta_u \) are obtained by expanding \( \alpha^{-1}(z) \beta(z) \) into a power series. Similarly, invertibility requires the roots of \( \beta(z) \) to be outside the unit circle.

The coefficients \( \theta_u \) of the representation \( X_t = \sum_{u=0}^{\infty} \theta_u \epsilon_{t-u} \) can also be computed recursively: setting \( \theta_k := 0 \) for \( k < 0 \), and \( \beta_0 := -1 \) as well as \( \beta_k := 0 \) for \( k > q \), it holds that

\[
\theta_k = \sum_{j=1}^{p} \alpha_j \theta_{k-j} - \beta_k. \tag{B.16}
\]

If \((X_t)_t\) is a weakly stationary and causal ARMA(p, q) process, then its mean is computed as

\[
\mu = \mathbb{E}[X_t] = \frac{1 - \beta_0}{1 - \alpha_0} \mu, \tag{B.17}
\]

while the autocorrelation function \( \rho(k) \) is obtained after solving the Yule–Walker equations

\[
\gamma(k) - \sum_{i=1}^{p} \alpha_i \gamma(|k-i|) = \begin{cases} 
-\sigma^2 \sum_{u=k}^{q} \beta_u \theta_{u-k} & \text{for } 0 \leq k < \max\{p, q + 1\}, \\
0 & \text{for } k \geq \max\{p, q + 1\}.
\end{cases} \tag{B.18}
\]

**B.4 Further Selected Models for Continuous-valued Time Series**

The basic ARMA models (which themselves might be applied to stationary processes with a short memory) gave rise to innumerable modifications and extensions to deal with, for example, trend or seasonality, long memory, time-varying volatility or with multivariate observations. A compact survey of time series models related to ARMA models is provided by Holan et al. (2010). Here, two of these models appear to be particularly relevant: the famous GARCH models with their ability to generate conditional heteroskedasticity, and the VARMA models as a multivariate extension of ARMA models.

**B.4.1 GARCH Models**

Let us start with the ARCH models, which were developed by R. F. Engle III in 1979. In 2003, he received (one half of) the Nobel prize in economic sciences “for methods of analyzing economic time series with time-varying volatility
An Introduction to Discrete-Valued Time Series

The ARCH models are motivated by a specific drawback of the AR models. If looking at the conditional mean and variance of an AR(p) process according to Definition B.3.3, then

\[ E[X_t \mid X_{t-1}, \ldots] = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \mu_e \]

varies in time according to the last p observations, while \( V[X_t \mid X_{t-1}, \ldots] = \sigma_e^2 \) is constant in time. But, especially for financial time series, it is common to observe clusters of large or low volatility, a phenomenon that cannot be reproduced by the AR models.

**Definition B.4.1.1 (ARCH(p) model)** Let \((\epsilon_t) \sim \mathcal{Z}\) be square-integrable white noise with \(E[\epsilon_t] = 0\) and \(V[\epsilon_t] = 1\). Then \((X_t) \sim \mathcal{Z}\) defined by

\[ X_t = \sigma_t \cdot \epsilon_t, \quad \text{where} \quad \sigma_t^2 = \beta_0 + \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2 \]

with \(\beta_0, \alpha_p > 0\) and \(\alpha_1, \ldots, \alpha_{p-1} \geq 0\), and where \(\epsilon_t\) is required to be independent of \((X_s)_{s < t}\) (causality), is said to be an autoregressive conditional heteroskedasticity process of order p, abbreviated as ARCH(p) process.

As a result, we obtain the time-varying conditional variances

\[ V[X_t \mid X_{t-1}, \ldots] = \sigma_t^2 = \beta_0 + \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2. \quad (B.19) \]

So now, the AR(p)-like recursion is not applied to the observations but to their conditional variances. In contrast, the unconditional variance remains constant in time provided that the requirement \(\sum_{j=1}^{p} \alpha_j < 1\) is satisfied. In fact, this condition (which is equivalent to requiring the roots of \(\alpha(z) := 1 - \alpha_1 z - \ldots - \alpha_p z^p\) to be outside the unit circle, due to the non-negativity of the \(\alpha_i\)) again guarantees the existence of a unique causal (weakly) stationary solution of the ARCH recursion in Definition B.4.1.1.

Looking at the autocorrelation structure, one might initially be surprised, since \(\gamma(k) = 0\) for \(k \neq 0\). So although an ARCH(p) process is obviously not serially independent by construction, it is serially uncorrelated. However, looking at the process of squared observations, autocorrelation becomes visible. If the weakly stationary and causal ARCH(p) process \((X_t) \sim \mathcal{Z}\) has existing fourth-order moments, then we can represent the squared process \((X_t^2) \sim \mathcal{Z}\) by an AR(p)-like recursion,

\[ X_t^2 = \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2 + \nu_t \]

with the \((\nu_t) \sim \mathcal{Z}\) being weak white noise having the mean \(E[\nu_t] = \beta_0\). Therefore, the autocorrelation function of the squared process satisfies the Yule–Walker equations (B.13) above.

---

2 A lot of background information concerning ARCH models can be found in Engle’s Nobel lecture at www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2003/.
A few years after the introduction of the ARCH model, Engle’s student T. Bollerslev proposed the generalized ARCH model, where the conditional variances not only depend on past observations but also on past conditional variances. So for a GARCH process of order \((p,q)\), abbreviated as \(GARCH(p,q)\) process, the recursion
\[
\sigma_t^2 = \beta_0 + \alpha_1 \sigma_{t-1}^2 + \ldots + \alpha_p \sigma_{t-p}^2 + \beta_1 \sigma_{t-1} + \ldots + \beta_q \sigma_{t-q}^2 \tag{B.20}
\]
is required to be satisfied. The condition for the existence of a (weakly) stationary and causal solution now becomes \(\alpha_1 + \ldots + \alpha_p + \beta_1 + \ldots + \beta_q < 1\).

### B.4.2 VARMA Models

As aforementioned, one half of the 2003’s Nobel prize in economic sciences was awarded to Engle; the other half was received by another statistician: C. W. J. Granger “for methods of analyzing economic time series with common trends (cointegration)," a topic related to multivariate time series. A multivariate (or vector-valued) time series \(x_1, \ldots, x_T\) is obtained if multiple features are observed simultaneously over time. For modeling the underlying multivariate (or vector) process \((X_t)_\mathcal{Z}\) with its possible cross-correlations, again a large variety of models is available; see Lütkepohl (2005) for a comprehensive survey. But in view of the scope of the present book, we shall only consider the vector autoregressive moving-average models of order \((p,q)\), abbreviated as \(VARMA(p,q)\) models.

Let \((X_t)_\mathcal{Z}\) be a process of \(d\)-dimensional random variables with range \(\mathbb{R}^d\). Then the mean vector \(\mu(t)\) and the autocovariance matrix \(\Gamma_t(k)\) are defined by
\[
\mu(t) := E[X_t] := (\ldots, E[X_{t,i}], \ldots)^T, \quad \Gamma_t(k) := E[(X_t - \mu(t)) (X_{t-k} - \mu(t-k))^T] := (E[(X_{t,i} - \mu_i(t)) (X_{t-k,j} - \mu_j(t-k))]_{i,j=1,\ldots,d}.
\] (B.21)

If both expressions exist (square-integrable) and if they are constant in time \(t\), then we refer to the process \((X_t)_\mathcal{Z}\) as being weakly stationary (as above), and we denote them as \(\mu\) and \(\Gamma(k)\) (without \(t\)) in this case. The autocorrelation matrix is then defined as
\[
R(k) = (\rho_{ij}(k))_{i,j=1,\ldots,d} \quad \text{with} \quad \rho_{ij}(k) = \frac{\gamma_{ij}(k)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}, \tag{B.22}
\]
where \(\gamma_{ij}(k)\) denotes the \((i,j)\)th element of \(\Gamma(k)\). Note that the definition (B.21) implies that
\[
\Gamma(k) = \Gamma(-k)^T. \tag{B.23}
\]

Analogous to Example B.1.2, the multivariate process \((\epsilon_t)_\mathcal{Z}\) is said to be a white noise if the \(\epsilon_t\) are i.i.d. If \((\epsilon_t)_\mathcal{Z}\) is also square-integrable with covariance matrix
An Introduction to Discrete-Valued Time Series

Σ_ε, then the corresponding autocovariance matrix equals

\[ \Gamma_ε(k) = \Sigma_ε \quad \text{for} \quad k = 0, \quad \text{and} \quad \Gamma_ε(k) = O \quad \text{otherwise}, \]  

(B.24)

with O being the zero matrix.

**Definition B.4.2.1 (VARMA(p, q) model)** Let \((ε_t)_z\) be square-integrable \(d\)-dimensional white noise. Then \((X_t)_z\) is said to be a **VARMA**(p, q) process if

\[ X_t = A_1 X_{t-1} + \ldots + A_p X_{t-p} + \varepsilon_t - B_1 \varepsilon_{t-1} - \ldots - B_q \varepsilon_{t-q} \]

with \(A_1, \ldots, B_q \in \mathbb{R}^{d \times d}\), or equivalently if

\[ A(B) X_t = B(B) \varepsilon_t \]

with \(A(z) = I - A_1 z - \ldots - A_p z^p\) and \(B(z) = I - B_1 z - \ldots - B_q z^q\), where I denotes the identity matrix.

The criterion for the existence of a unique causal and stationary solution now becomes

\[ \det(A(z)) \neq 0 \quad \text{for} \quad |z| \leq 1, \]

(B.25)

while the criterion for invertibility is given by

\[ \det(B(z)) \neq 0 \quad \text{for} \quad |z| \leq 1. \]

(B.26)

The autocovariance matrices of a causal stationary VARMA(p, q) process are determined again by a set of Yule–Walker equations (by considering (B.23)), given by

\[ \Gamma(k) - \sum_{i=1}^{p} A_i \Gamma(k-i) = \sum_{j=k}^{q} B_j \Sigma_\epsilon C_{j-k}^T \quad \text{for} \quad k = 0, 1, \ldots, \]

(B.27)

where \(C_r = O\) for \(r < 0\), \(C_0 = I\)

and \(C_r = \sum_{i=1}^{r} A_i C_{r-i} + B_r\) for \(r > 0\).

Up to this point, everything looks analogous to the corresponding results for the univariate ARMA(p, q) models; see Definition B.3.4 and below. Now, we have to discuss some important differences. First, it has to be mentioned that the VAR(1) model is of outstanding importance, since any \(d\)-dimensional VARMA(p, q) model can be translated into an \(d(p + q)\)-dimensional VAR(1) model. This is done in an analogous way to the approach for Markov processes sketched below (B.1), by constructing the coefficient matrix \(D \in \mathbb{R}^{d(p+q) \times d(p+q)}\) as follows:

\[ D_{11} = \begin{pmatrix}
A_1 & \ldots & A_{p-1} & A_p \\
I & O & O & \\
& \ddots & \vdots & \\
O & I & O
\end{pmatrix} \in \mathbb{R}^{dp \times dp}, \]

(B.28)
\[
D_{12} = \begin{pmatrix}
B_1 & \ldots & B_q \\
O & \ldots & O \\
\vdots & \ddots & \vdots \\
O & \ldots & O \\
\end{pmatrix} \in \mathbb{R}^{dp \times dq}, \quad D_{21} = O \in \mathbb{R}^{dq \times dp},
\]
\[
D_{22} = \begin{pmatrix}
O & \ldots & O & O \\
I & O & O & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
O & I & O & O \\
\end{pmatrix} \in \mathbb{R}^{dq \times dq}, \quad D = \begin{pmatrix} D_{11} & D_{12} \\
D_{21} & D_{22} \end{pmatrix}.
\]

With
\[
Y_t := \begin{pmatrix}
X_t \\
\vdots \\
X_{t-p+1} \\
\epsilon_t \\
\vdots \\
\epsilon_{t-q+1}
\end{pmatrix}, \quad U_t := \begin{pmatrix}
\epsilon_t \\
0 \\
\vdots \\
0 \\
\epsilon_t \\
\vdots \\
0
\end{pmatrix},
\]
we have
\[
Y_t = D Y_{t-1} + U_t. \tag{B.29}
\]

Using this VAR(1)-representation, we may, for example, trace back the autocorrelation function of the general VARMA(p, q) model to that of the VAR(1) model. The latter is given by
\[
\Gamma(k) = A_1^k \Gamma(0), \quad \Gamma(0) = A_1 \Gamma(0) A_1^\top + \Sigma_e, \tag{B.30}
\]
see (B.27) with \((p, q) = (1, 0)\). Furthermore, the causal solution of the VAR(1) model equals
\[
X_t = \sum_{j=0}^{\infty} A_1^j \epsilon_{t-j}. \tag{B.31}
\]

As the second main difference to the univariate ARMA models, it has to be mentioned that sometimes there is not a unique representation of a VARMA process (counterexamples are easily constructed; see for example the one on p. 259 in Holan et al. (2010)). Therefore, in practice, one often restricts oneself to the purely autoregressive VAR(p) models in advance to avoid such non-identifiability issues. Further information is presented in the book by Lütkepohl (2005).